(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103049433 A
(43)申请公布日 2013.04.17

(21)申请号 201210533328.0

(22)申请日 2012.12.11

(71)申请人 微梦创科网络科技(中国)有限公司
地址 100080 北京市海淀区海淀北二街 10 号 701 室

(72)发明人 陈开江

(74)专利代理机构 北京市京大律师事务所 11321
代理人 黄启行 方晓明

(51)Int.Cl.
G06F 17/27(2006.01)
G06F 17/30(2006.01)

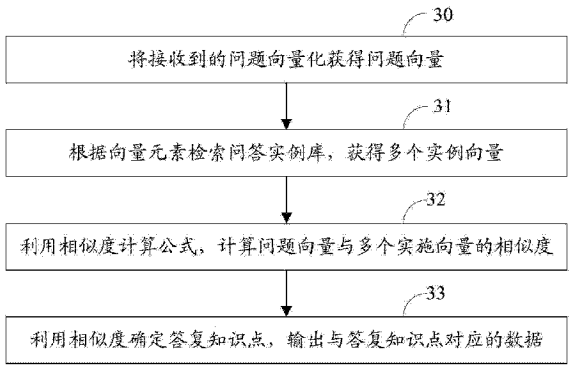
权利要求书 5 页 说明书 13 页 附图 3 页

(54)发明名称

自动问答方法、自动问答系统及构建问答实例库的方法

(57)摘要

本申请公开了一种自动问答方法、自动问答系统及构建问答实例库的方法。具体地，利用以向量方式存储问答实例的问答实例库，将用户提交的问题向量化生成问题向量，利用问题向量及包含实例向量的问答实例库，查找相似度符合要求的回答知识点 ID，再利用回答知识点 ID 从知识点文案库中获取输出给用户的答案内容。采用本发明的系统及方法，能够降低成本，提高工作效率。



CN 103049433 A

1. 一种自动问答方法，其特征在于，该方法包括：

A、将接收到的问题向量化获得问题向量；所述问题向量包含多个向量元素；

B、根据所述向量元素检索问答实例库，获得多个实例向量；任一所述实例向量至少包含一个向量元素；

C、利用相似度计算公式，计算问题向量与多个实例向量的相似度；

D、利用所述相似度确定答复知识点，输出与答复知识点对应的数据。

2. 根据权利要求 1 所述的方法，其特征在于，所述步骤 A 之前进一步包括：

A'、采样人工回答记录并向量化，生成问答实例库。

3. 根据权利要求 2 所述的方法，其特征在于，所述步骤 A' 包括：

A'_1、确定需自动问答的知识点，为所述需自动问答的知识点分配问题 ID；

A'_2、根据所述需自动问答的知识点，对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例，为所述问答实例包含的知识点分配回答知识点 ID；

A'_3、向量化所述问答实例包含的问题，获得问题向量；

A'_4、将所述问答实例以三元组的形式进行存储；任一所述问答实例的三元组包含问题 ID、问题向量及回答知识点 ID。

4. 根据权利要求 3 所述的方法，其特征在于，步骤 A'_2 所述对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例包括：

A'_{21}、确定进行采样的知识点的样本数量 n；所述进行采样的知识点为需自动回答的知识点；所述 n 为自然数；

A'_{22}、从人工回答记录中选择包含所述知识点的 n 个问题实例；

A'_{23}、计算所述 n 个问题实例的问题的整体长度方差；

A'_{24}、判断所述问题的整体长度方差是否低于第一阈值，如果是，则执行步骤 A'_{25}，否则，去除所述 n 个问题实例中问题长度与所述 n 个问题实例的长度平均值的差值最大的一个问题实例，从人工回答记录中再选择一个包含所述知识点的问题实例，执行步骤 A'_{23}；

A'_{25}、将所述 n 个问题实例作为选择的 n 个包含所述需自动回答的知识点的问答实例。

5. 根据权利要求 4 所述的方法，其特征在于，所述步骤 A'_{23} 为：

利用$\frac{1}{n}\sum_{q_i \in K_j}\left(len(q_i)-E_{K_j}(len)\right)^2$计算所述 n 个问题实例的问题的整体长度方差；

所述 q_i 为知识点 K_j 的问题样本，所述 $len(q_i)$ 为 q_i 包含的词数量，所述$E_{K_j}(len)$为知识点 K_j 中所有问题长度的平均值。

6. 根据权利要求 3 所述的方法，其特征在于，所述步骤 A'_3 包括：

提取所述问答实例中问题的关键词、二元字符串及特殊词性；

将所述关键词作为向量元素，计算每一向量元素的权重；

利用每个向量元素的权重计算问答实例的向量的长度；

将向量元素、向量元素的权重及向量的长度作为问题向量。

7. 根据权利要求 6 所述的方法，其特征在于，所述计算每一向量元素的权重包括：

利用$ch(i)=\frac{N*[A*N-CF*TF]^2}{CF*TF*(N-CF)*(N-TF)}$计算获得向量元素与知识点之间的卡方值 ch(i)；

利用 weight(c,a)=ln(chi+b) 计算获得预设的知识点范围内的每一个向量元素的权重；

所述 N 为样本总数量，所述 CF 为每个知识点的样本数量，所述 TF 为每个向量元素出现的样本数量，所述 A 为向量元素和知识点共同出现的样本数量，所述 ch(i) 为向量元素 c 与知识点 a 的卡方值；

所述 weight(c,a) 表示向量元素 c 在知识点 a 中的权重，b 是平滑值，所述 b 为小数。

8. 根据权利要求 6 所述的方法，其特征在于，所述利用每个向量元素的权重计算问答实例的向量的长度包括：

利用$|v|=\sqrt{\sum_{i}^{n} w_i * w_i}$计算问答实例的向量长度；

所述 |v| 为向量长度，所述 w_i 为向量问答实例中每个向量元素的权重。

9. 根据权利要求 1-8 任一项所述的方法，其特征在于，所述步骤 A 包括：

将接收到的问题规整、分词和归一化处理，提取关键词、二元字符串和词性；

将关键词作为向量元素；

将向量元素、向量元素的权重及向量的长度作为问题向量；所述向量元素的权重和所述向量长度为空。

10. 根据权利要求 1-8 任一项所述的方法，其特征在于，所述步骤 C 包括：

C1、将所述多个实例向量按照其包含的回答知识点 ID 进行归类；

C2、对于同一回答知识点 ID，动态调整问题向量包含的向量元素在相同的回答知识点 ID 对应的实例向量中的权重，获得调整后的向量元素；

C3、利用所述调整后的向量元素在所述相同回答知识点 ID 对应的实例向量中的权重，计算所述调整后的向量元素与所述实例向量的余弦相似度。

11. 根据权利要求 10 所述的方法，其特征在于，所述步骤 C3 包括：

利用$sim(v_q, v_c)=\frac{\sum_{i} w_i^{(q)} * w_i^{(c)}}{|v_q| * |v_c|}$计算所述调整后的向量元素所在的问题向量与所述实例向量的余弦相似度；

所述 v_q 为问题向量；所述 v_c 为实例向量；所述 $|v_q|*|v_c|$ 表示问题向量的长度与实例向量的长度的乘积；所述$w_i^{(q)}$和$w_i^{(c)}$表示两个向量中相同向量元素对应的权重。

12. 根据权利要求 1-8 任一项所述的方法，其特征在于，所述步骤 D 包括：

D1、将所述相似度转换为实例向量与问题向量之间的距离；

D2、将所述距离小于第二阈值的实例向量作为候选实例向量；

D3、利用候选实例向量计算获得其对应的回答知识点 ID 的加权票数；

D4、在回答知识点 ID 的平均票数大于第三阈值时或在为回答知识点 ID 投票的实例向量的数量大于第四阈值时，将回答知识点 ID 确定为候选知识点 ID；

D5、选择加权票数排列在前 L 位的候选知识点 ID 对应的知识点为答复知识点；所述 L 为自然数；

D6、利用候选知识点 ID 从知识点文案库中读取排列在前 L 位的回答内容，并按照加权票数的排列顺序输出 L 个回答内容。

13. 根据权利要求 12 所述的方法，其特征在于，所述步骤 D1 包括：

利用$dis(v_c, v_q) = \log(\frac{1}{sim(v_c, v_q)})$及 $sim(v_c, v_q) > \lambda > 0$ 计算实例向量与问题向量之间的距离；

所述 $sim(v_c, v_q)$ 为所述实例向量与所述问题向量的余弦相似度；

所述 λ 为相似度阈值。

14. 根据权利要求 12 所述的方法，其特征在于，所述步骤 D3 包括：

利用$vote(ID_j) = \sum_{v_i \in ID_j}^{m} \frac{1}{dis < v_i, v_q >^2}$计算候选实例向量对应的回答知识点 ID 的加权票数；

所述 ID_j 为实例向量包含的回答知识点 ID；所述 v_i 为属于回答知识点 ID_j 的实例向量；所述 v_q 是问题向量；所述 $dis < v_i, v_q >$ 为实例向量与问题向量之间的距离；所述 m 为属于回答知识点 ID_j 的实例向量的数量。

15. 根据权利要求 12 所述的方法，其特征在于，步骤 D4 所述回答知识点 ID 的平均票数为回答知识点 ID 所述加权票数除以属于回答知识点 ID 的实例向量的数量获得的商。

16. 一种自动问答系统，其特征在于，该系统包含：

问答实例库，以三元组的形式存储问答实例；任一所述问答实例的三元组包含问题 ID、问题向量及回答知识点 ID；

问题解析模块，将接收到的问题向量化获得问题向量；所述问题向量包含多个向量元素；

所述问题解析单元根据所述向量元素检索所述问答实例库，获得多个实例向量，并输出至答案生成模块；任一所述实例向量至少包含一个向量元素；

答案生成模块，利用相似度计算公式，计算问题向量与多个实例向量的相似度，利用所述相似度确定答复知识点，输出与答复知识点对应的数据。

17. 根据权利要求 16 所述的系统，其特征在于，该系统还包含：

构建模块，采样人工回答记录并向量化，生成问答实例库。

18. 根据权利要求 16 或 17 所述的系统，其特征在于，所述问题解析模块包含：

第一向量化单元，将接收到的问题规整、分词和归一化处理，提取关键词、二元字符串和词性，将关键词作为向量元素，将向量元素、向量元素的权重及向量的长度作为问题向量；所述向量元素的权重和所述向量长度为空；

检索单元，根据所述向量元素检索所述问答实例库，获得多个实例向量，并输出至所述答案生成模块。

19. 根据权利要求 16 或 17 所述的系统，其特征在于，所述答案生成模块包括：

实例挑选单元，将所述多个实例向量按照其包含的回答知识点 ID 进行归类，对于同一回答知识点 ID，动态调整问题向量包含的向量元素在相同的回答知识点 ID 对应的实例向量中的权重，获得调整后的向量元素；

所述实例挑选单元利用所述调整后的向量元素在所述相同回答知识点 ID 对应的实例向量中的权重，计算所述调整后的向量元素与所述实例向量的距离，将所述距离小于第二阈值的实例向量作为候选实例向量并输出至知识点挑选单元；

知识点挑选单元，利用候选实例向量计算获得其对应的回答知识点 ID 的加权票数，在回答知识点 ID 的平均票数大于第三阈值时或在为回答知识点 ID 投票的实例向量的数量大于第四阈值时，将回答知识点 ID 确定为候选知识点 ID 并输出至答案筛选单元；

答案筛选单元，选择加权票数排列在前 L 位的候选知识点 ID 对应的知识点为答复知识点，利用候选知识点 ID 从知识点文案库中读取排列在前 L 位的回答内容，并按照加权票数的排列顺序输出 L 个回答内容；所述 L 为自然数。

20. 根据权利要求 17 所述的系统，其特征在于，所述构建模块包含：

知识点确定单元，确定需自动问答的知识点，为所述需自动问答的知识点分配问题 ID，输出所述需自动问答的知识点及其对应的所述问题 ID 至采样单元；

采样单元，根据所述需自动问答的知识点，对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例，为所述问答实例包含的回答数据分配回答知识点 ID，输出问题 ID、回答知识点 ID 及问答实例至第二向量化单元；

第二向量化单元，向量化所述问答实例包含的问题，获得问题向量，将所述问答实例以三元组的形式存储于问答实例库中；任一所述问答实例的三元组包含问题 ID、问题向量及回答知识点 ID。

21. 根据权利要求 16 或 17 所述的系统，其特征在于，该系统还包含：

知识点文案库，以三元组形式保存知识点向量；任一所述知识点向量的三元组包含回答知识点 ID、知识点描述及知识点回答文案。

22. 一种构建问答实例库的方法，其特征在于，该方法包括：

A、确定需自动问答的知识点，为所述需自动问答的知识点分配问题 ID；

B、根据所述需自动问答的知识点，对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例，为所述问答实例包含的知识点分配回答知识点 ID；

C、向量化所述问答实例包含的问题，获得问题向量；

D、将所述问答实例以三元组的形式进行存储；任一所述问答实例的三元组包含问题 ID、问题向量及回答知识点 ID。

23. 根据权利要求 22 所述的方法，其特征在于，步骤 B 所述对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例包括：

B1、确定进行采样的知识点的样本数量 n；所述进行采样的知识点为需自动回答的知识点；所述 n 为自然数；

B2、从人工回答记录中选择包含所述知识点的 n 个问题实例；

B3、计算所述 n 个问题实例的问题的整体长度方差；

B4、判断所述问题的整体长度方差是否低于第一阈值，如果是，则执行步骤 B5，否则，去除所述 n 个问题实例中问题长度与所述 n 个问题实例的长度平均值的差值最大的一个问题实例，从人工回答记录中再选择一个包含所述知识点的问题实例，执行步骤 B3；

B5、将所述 n 个问题实例作为选择的 n 个包含所述需自动回答的知识点的问答实例。

24. 根据权利要求 23 所述的方法，其特征在于，所述步骤 B3 为：

利用$\frac{1}{n}\sum_{q_i \in K_j}\left(len(q_i)-E_{K_j}(len)\right)^2$计算所述 n 个问题实例的问题的整体长度方差；

所述 q_i 为知识点 K_j 的问题样本，所述 $len(q_i)$ 为 q_i 包含的词数量，所述$E_{K_j}(len)$为知识

点 K_j 中所有问题长度的平均值。

25. 根据权利要求 22 所述的方法，其特征在于，所述步骤 C 包括：

C1、提取所述问答实例中问题的关键词、二元字符串及特殊词性；

C2、将所述关键词作为向量元素，计算每一向量元素的权重；

C3、利用每个向量元素的权重计算问答实例的向量的长度；

C4、将向量元素、向量元素的权重及向量的长度作为问题向量。

26. 根据权利要求 25 所述的方法，其特征在于，所述步骤 C2 包括：

利用 $ch(i)=\frac{N*[A*N-CF*TF]^2}{CF*TF*(N-CF)*(N-TF)}$ 计算获得向量元素与知识点之间的卡方值 ch(i)；

利用 weight(c,a)=ln(chi+b) 计算获得预设的知识点范围内的每一个向量元素的权重；

所述 N 为样本总数量，所述 CF 为每个知识点的样本数量，所述 TF 为每个向量元素出现的样本数量，所述 A 为向量元素和知识点共同出现的样本数量，所述 ch(i) 为向量元素 c 与知识点 a 的卡方值；

所述 weight(c,a) 表示向量元素 c 在知识点 a 中的权重，b 是平滑值，所述 b 为小数。

27. 根据权利要求 25 所述的方法，其特征在于，所述步骤 C3 包括：利用 $|v|=\sqrt{\sum_i^n w_i*w_i}$ 计算问答实例的向量长度；

所述 |v| 为向量长度，所述 w_i 为向量问答实例中每个向量元素的权重。

自动问答方法、自动问答系统及构建问答实例库的方法

技术领域

[0001] 本发明涉及计算机自然语言处理领域，特别涉及一种自动问答方法、一种自动问答系统及一种构建问答实例库的方法。

背景技术

[0002] 目前，很多行业需要承担越来越多的用户咨询和反馈的解答工作，比如互联网行业的售后服务或者客户服务。由于用户数量的指数增长，已经无法采用人工的方式对所有用户的咨询进行反馈或及时回答，并且用户的问题大多集中在某些特定的知识点上，人工回复往往是进行重复性地劳动，因此，急需一种简单、高效、易维护的系统来辅助人工进行问题回复。

[0003] 自动问答(Question Answering，QA)是指根据用户的自然语言提出的问题找到一个明确的答案。图 1 为现有的自动问答系统的结构示意图，现结合图 1，对现有的自动问答系统的结构进行说明，具体如下：

[0004] 现有的自动问答系统包括：接口单元 101、推理单元 102 和知识库 103。接口单元 101 将用户采用自然语言进行提问的问题发送给推理单元 102，推理单元 102 对问题进行解析得到问题的结构化表达及关键词，根据问题的结构化表达式及关键词从知识库 103 中匹配获得相关的应答内容，利用问题的结构化表达式、本体知识技术及语言知识技术从知识库 103 中匹配获得问题模板，利用自然语言处理技术、获得的应答内容及获得的问题模板，完成知识推理并最终生成答案，通过接口单元 101 输出生成的答案。

[0005] 现有的自动问答系统的知识库 103 的构建阶段，需要从输入的新问答对中挖掘问题模板，构建出模板库，以供推理单元 102 查询获得问题模板，模板库中的问题模板可为句型模板、语义模板等；可采用语言知识技术对输入的新知识进行处理以获得与关键词对应的应答，还可对输入的新问答对进行知识解析以获得与关键词对应的应答。知识库 103 中保存的知识即为与关键词对应的应答，并且知识库 103 采用人工智能领域常用的本体知识表示知识，而上述知识库 103 的构建都需要人工完成。

[0006] 现有的自动问答系统的知识库构建和维护成本较大，且需要靠业务人员总结某个知识点的常见问题模板，且运维需要持续加入新模板；由于模板库会越来越大，推理单元进行推理计算的过程会越来越耗时，造成计算复杂，响应时间不可控，工作效率低下，现有的自动问答系统还有待进一步改进。

发明内容

[0007] 本发明提供了一种自动问答方法及系统，用以使得的项目或应用得以实现。

[0008] 根据本发明的一个方面，提供了一种自动问答方法，该方法能够降低成本，提高工作效率。

[0009] 根据本发明的一个方面，提供了一种自动问答系统，该系统能够降低成本，提高工作效率。

[0010] 根据本发明的一个方面，提供了一种构建问答实例库的方法。

[0011] 本发明的目的是通过下述技术方案实现的：

[0012] 本发明提供了一种自动问答方法，该方法包括：

[0013] A、将接收到的问题向量化获得问题向量；所述问题向量包含多个向量元素；

[0014] B、根据所述向量元素检索问答实例库，获得多个实例向量；任一所述实例向量至少包含一个向量元素；

[0015] C、利用相似度计算公式，计算问题向量与多个实例向量的相似度；

[0016] D、利用所述相似度确定答复知识点，输出与答复知识点对应的数据。

[0017] 较佳地，所述步骤 A 之前进一步包括：

[0018] A’、采样人工回答记录并向量化，生成问答实例库。

[0019] 上述方法中，所述步骤 A’ 包括：

[0020] A'_1、确定需自动问答的知识点，为所述需自动问答的知识点分配问题 ID；

[0021] A'_2、根据所述需自动问答的知识点，对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例，为所述问答实例包含的知识点分配回答知识点 ID；

[0022] A'_3、向量化所述问答实例包含的问题，获得问题向量；

[0023] A'_4、将所述问答实例以三元组的形式进行存储；任一所述问答实例的三元组包含问题 ID、问题向量及回答知识点 ID。

[0024] 上述方法中，步骤 A'_2 所述对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例包括：

[0025] A'_{21}、确定进行采样的知识点的样本数量 n；所述进行采样的知识点为需自动回答的知识点；所述 n 为自然数；

[0026] A'_{22}、从人工回答记录中选择包含所述知识点的 n 个问题实例；

[0027] A'_{23}、计算所述 n 个问题实例的问题的整体长度方差；

[0028] A'_{24}、判断所述问题的整体长度方差是否低于第一阈值，如果是，则执行步骤 A'_{25}，否则，去除所述 n 个问题实例中问题长度与所述 n 个问题实例的长度平均值的差值最大的一个问题实例，从人工回答记录中再选择一个包含所述知识点的问题实例，执行步骤 A'_{23}；

[0029] A'_{25}、将所述 n 个问题实例作为选择的 n 个包含所述需自动回答的知识点的问答实例。

[0030] 上述方法中，所述步骤 A'_{23} 为：

[0031] 利用$\frac{1}{n}\sum_{q_i \in K_j}\left(len(q_i)-E_{K_j}(len)\right)^2$计算所述 n 个问题实例的问题的整体长度方差；

[0032] 所述 q_i 为知识点 K_j 的问题样本，所述 $len(q_i)$ 为 q_i 包含的词数量，所述$E_{K_j}(len)$为知识点 K_j 中所有问题长度的平均值。

[0033] 上述方法中，所述步骤 A'_3 包括：

[0034] 提取所述问答实例中问题的关键词、二元字符串及特殊词性；

[0035] 将所述关键词作为向量元素，计算每一向量元素的权重；

[0036] 利用每个向量元素的权重计算问答实例的向量的长度；

[0037] 将向量元素、向量元素的权重及向量的长度作为问题向量。

[0038] 上述方法中，所述计算每一向量元素的权重包括：

[0039] 利用$ch(i)=\frac{N*[A*N-CF*TF]^2}{CF*TF*(N-CF)*(N-TF)}$计算获得向量元素与知识点之间的卡方值 ch(i)；

[0040] 利用 weight(c,a)=ln(chi+b) 计算获得预设的知识点范围内的每一个向量元素的权重；

[0041] 所述 N 为样本总数量，所述 CF 为每个知识点的样本数量，所述 TF 为每个向量元素出现的样本数量，所述 A 为向量元素和知识点共同出现的样本数量，所述 ch(i) 为向量元素 c 与知识点 a 的卡方值；

[0042] 所述 weight(c,a) 表示向量元素 c 在知识点 a 中的权重，b 是平滑值，所述 b 为小数。

[0043] 上述方法中，所述利用每个向量元素的权重计算问答实例的向量的长度包括：

[0044] 利用$|v|=\sqrt{\sum_i^n w_i*w_i}$计算问答实例的向量长度；

[0045] 所述 |v| 为向量长度，所述 w_i 为向量问答实例中每个向量元素的权重。

[0046] 上述方法中，所述步骤 A 包括：

[0047] 将接收到的问题规整、分词和归一化处理，提取关键词、二元字符串和词性；

[0048] 将关键词作为向量元素；

[0049] 将向量元素、向量元素的权重及向量的长度作为问题向量；所述向量元素的权重和所述向量长度为空。

[0050] 上述方法中，所述步骤 C 包括：

[0051] C1、将所述多个实例向量按照其包含的回答知识点 ID 进行归类；

[0052] C2、对于同一回答知识点 ID，动态调整问题向量包含的向量元素在相同的回答知识点 ID 对应的实例向量中的权重，获得调整后的向量元素；

[0053] C3、利用所述调整后的向量元素在所述相同回答知识点 ID 对应的实例向量中的权重，计算所述调整后的向量元素与所述实例向量的余弦相似度。

[0054] 上述方法中，所述步骤 C3 包括：

[0055] 利用$sim(v_q,v_c)=\frac{\sum_i w_i^{(q)}*w_i^{(c)}}{|v_q|*|v_c|}$计算所述调整后的向量元素所在的问题向量与所述实例向量的余弦相似度；

[0056] 所述 v_q 为问题向量；所述 v_c 为实例向量；所述 $|v_q|*|v_c|$ 表示问题向量的长度与实例向量的长度的乘积；所述$w_i^{(q)}$和$w_i^{(c)}$表示两个向量中相同向量元素对应的权重。

[0057] 上述方法中，所述步骤 D 包括：

[0058] D1、将所述相似度转换为实例向量与问题向量之间的距离；

[0059] D2、将所述距离小于第二阈值的实例向量作为候选实例向量；

[0060] D3、利用候选实例向量计算获得其对应的回答知识点 ID 的加权票数；

[0061] D4、在回答知识点 ID 的平均票数大于第三阈值时或在为回答知识点 ID 投票的实例向量的数量大于第四阈值时，将回答知识点 ID 确定为候选知识点 ID；

[0062] D5、选择加权票数排列在前 L 位的候选知识点 ID 对应的知识点为答复知识点；所

述 L 为自然数；

[0063] D6、利用候选知识点 ID 从知识点文案库中读取排列在前 L 位的回答内容，并按照加权票数的排列顺序输出 L 个回答内容。

[0064] 上述方法中，所述步骤 D1 包括：

[0065] 利用$dis(v_c, v_q) = \log(\frac{1}{sim(v_c, v_q)})$及 $sim(v_c, v_q) > \lambda > 0$ 计算实例向量与问题向量之间的距离；

[0066] 所述 $sim(v_c, v_q)$ 为所述实例向量与所述问题向量的余弦相似度；

[0067] 所述 λ 为相似度阈值。

[0068] 上述方法中，所述步骤 D3 包括：

[0069] 利用$vote(ID_j) = \sum_{v_i \in ID_j}^{m} \frac{1}{dis < v_i, v_q >^2}$计算候选实例向量对应的回答知识点 ID 的加权票数；

[0070] 所述 ID_j 为实例向量包含的回答知识点 ID；所述 v_i 为属于回答知识点 ID_j 的实例向量；所述 v_q 是问题向量；所述 $dis < v_i, v_q >$ 为实例向量与问题向量之间的距离；所述 m 为属于回答知识点 ID_j 的实例向量的数量。

[0071] 上述方法中，步骤 D4 所述回答知识点 ID 的平均票数为回答知识点 ID 所述加权票数除以属于回答知识点 ID 的实例向量的数量获得的商。

[0072] 本发明提供了一种自动问答系统，该系统包含：

[0073] 问答实例库，以三元组的形式存储问答实例；任一所述问答实例的三元组包含问题 ID、问题向量及回答知识点 ID；

[0074] 问题解析模块，将接收到的问题向量化获得问题向量；所述问题向量包含多个向量元素；

[0075] 所述问题解析单元根据所述向量元素检索所述问答实例库，获得多个实例向量，并输出至答案生成模块；任一所述实例向量至少包含一个向量元素；

[0076] 答案生成模块，利用相似度计算公式，计算问题向量与多个实例向量的相似度，利用所述相似度确定答复知识点，输出与答复知识点对应的数据。

[0077] 较佳地，该系统还包含：

[0078] 构建模块，采样人工回答记录并向量化，生成问答实例库。

[0079] 上述系统中，所述问题解析模块包含：

[0080] 第一向量化单元，将接收到的问题规整、分词和归一化处理，提取关键词、二元字符串和词性，将关键词作为向量元素，将向量元素、向量元素的权重及向量的长度作为问题向量；所述向量元素的权重和所述向量长度为空；

[0081] 检索单元，根据所述向量元素检索所述问答实例库，获得多个实例向量，并输出至所述答案生成模块。

[0082] 上述系统中，所述答案生成模块包括：

[0083] 实例挑选单元，将所述多个实例向量按照其包含的回答知识点 ID 进行归类，对于同一回答知识点 ID，动态调整问题向量包含的向量元素在相同的回答知识点 ID 对应的实例向量中的权重，获得调整后的向量元素；

[0084] 所述实例挑选单元利用所述调整后的向量元素在所述相同回答知识点 ID 对应的实例向量中的权重，计算所述调整后的向量元素与所述实例向量的距离，将所述距离小于第二阈值的实例向量作为候选实例向量并输出至知识点挑选单元；

[0085] 知识点挑选单元，利用候选实例向量计算获得其对应的回答知识点 ID 的加权票数，在回答知识点 ID 的平均票数大于第三阈值时或在为回答知识点 ID 投票的实例向量的数量大于第四阈值时，将回答知识点 ID 确定为候选知识点 ID 并输出至答案筛选单元；

[0086] 答案筛选单元，选择加权票数排列在前 L 位的候选知识点 ID 对应的知识点为答复知识点，利用候选知识点 ID 从知识点文案库中读取排列在前 L 位的回答内容，并按照加权票数的排列顺序输出 L 个回答内容；所述 L 为自然数。

[0087] 上述系统中，所述构建模块包含：

[0088] 知识点确定单元，确定需自动问答的知识点，为所述需自动问答的知识点分配问题 ID，输出所述需自动问答的知识点及其对应的所述问题 ID 至采样单元；

[0089] 采样单元，根据所述需自动问答的知识点，对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例，为所述问答实例包含的回答数据分配回答知识点 ID，输出问题 ID、回答知识点 ID 及问答实例至第二向量化单元；

[0090] 第二向量化单元，向量化所述问答实例包含的问题，获得问题向量，将所述问答实例以三元组的形式存储于问答实例库中；任一所述问答实例的三元组包含问题 ID、问题向量及回答知识点 ID。

[0091] 较佳地，该系统还包含：

[0092] 知识点文案库，以三元组形式保存知识点向量；任一所述知识点向量的三元组包含回答知识点 ID、知识点描述及知识点回答文案。

[0093] 本发明提供了一种构建问答实例库的方法，该方法包括：

[0094] A、确定需自动问答的知识点，为所述需自动问答的知识点分配问题 ID；

[0095] B、根据所述需自动问答的知识点，对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例，为所述问答实例包含的知识点分配回答知识点 ID；

[0096] C、向量化所述问答实例包含的问题，获得问题向量；

[0097] D、将所述问答实例以三元组的形式进行存储；任一所述问答实例的三元组包含问题 ID、问题向量及回答知识点 ID。

[0098] 上述方法中，步骤 B 所述对人工回答记录进行采样，获得与所述需自动问答的知识点对应的问答实例包括：

[0099] B1、确定进行采样的知识点的样本数量 n；所述进行采样的知识点为需自动回答的知识点；所述 n 为自然数；

[0100] B2、从人工回答记录中选择包含所述知识点的 n 个问题实例；

[0101] B3、计算所述 n 个问题实例的问题的整体长度方差；

[0102] B4、判断所述问题的整体长度方差是否低于第一阈值，如果是，则执行步骤 B5，否则，去除所述 n 个问题实例中问题长度与所述 n 个问题实例的长度平均值的差值最大的一个问题实例，从人工回答记录中再选择一个包含所述知识点的问题实例，执行步骤 B3；

[0103] B5、将所述 n 个问题实例作为选择的 n 个包含所述需自动回答的知识点的问答实例。

[0104] 上述方法中，所述步骤 B3 为：

[0105] 利用$\frac{1}{n}\sum_{q_i \in K_j}\left(len(q_i)-E_{K_j}(len)\right)^2$计算所述 n 个问题实例的问题的整体长度方差；

[0106] 所述 q_i 为知识点 K_j 的问题样本，所述 $len(q_i)$ 为 q_i 包含的词数量，所述$E_{K_j}(len)$为知识点 K_j 中所有问题长度的平均值。

[0107] 上述方法中，所述步骤 C 包括：

[0108] C1、提取所述问答实例中问题的关键词、二元字符串及特殊词性；

[0109] C2、将所述关键词作为向量元素，计算每一向量元素的权重；

[0110] C3、利用每个向量元素的权重计算问答实例的向量的长度；

[0111] C4、将向量元素、向量元素的权重及向量的长度作为问题向量。

[0112] 上述方法中，所述步骤 C2 包括：

[0113] 利用$ch(i)=\frac{N*[A*N-CF*TF]^2}{CF*TF*(N-CF)*(N-TF)}$计算获得向量元素与知识点之间的卡方值 ch(i)；

[0114] 利用 weight(c,a)=ln(chi+b) 计算获得每一个向量元素的权重；

[0115] 所述 N 为样本总数量，所述 CF 为每个知识点的样本数量，所述 TF 为每个向量元素出现的样本数量，所述 A 为向量元素和知识点共同出现的样本数量，所述 ch(i) 为向量元素 c 与知识点 a 的卡方值；

[0116] 所述 weight(c,a) 表示向量元素 c 在知识点 a 中的权重，b 是平滑值，所述 b 为小数。

[0117] 上述方法中，所述步骤 C3 包括：

[0118] 利用$|v|=\sqrt{\sum_i^n w_i * w_i}$计算问答实例的向量长度；

[0119] 所述 |v| 为向量长度，所述 w_i 为向量问答实例中每个向量元素的权重。

[0120] 由上述的技术方案可见，本发明提供了一种自动问答方法及系统，利用以向量方式存储问答实例的问答实例库，将用户提交的问题向量化生成问题向量，利用问题向量及包含实例向量的问答实例库，查找相似度符合要求的回答知识点 ID，再利用回答知识点 ID 从知识点文案库中获取输出给用户的答案内容。本发明还提供了一种构建问答实例库的方法。采用本发明的系统及方法，能够降低成本，提高工作效率。

附图说明

[0121] 图 1 为现有的自动问答系统的结构示意图；

[0122] 图 2 为本发明构建问答实例库的方法流程图；

[0123] 图 3 为本发明自动问答方法的流程图；

[0124] 图 4 为本发明实例向量的分类示意图；

[0125] 图 5 为本发明自动问答系统的结构示意图。

具体实施方式

[0126] 由于现有技术中采用模板匹配的方法实现自动问答，知识库构建和维护成本较

大，随着模板数量的持续性增长，匹配模板以获得答案的工作效率逐步降低，而本发明的自动问答方法中，对人工问答记录进行处理生成以向量方式存储问答实例的问答实例库，将用户提交的问题向量化生成问题向量，利用问题向量及包含实例向量的问答实例库，查找符合要求的回答知识点 ID，再利用回答知识点 ID 从知识点文案库中获取输出给用户的答案内容，不仅系统的运维成本较低，而且整个自动问答过程的工作效率得到了显著地提升。

[0127] 为了表述清楚，先对本发明涉及的专业词汇进行说明，具体如下：

[0128] 知识点就是用户诉求（包括咨询、反馈等）的话题，比如围绕“如何修改登录密码”这个话题，用户会以各种不同的表述方式表达其诉求即用户实际的问题，那么这个话题就是一个知识点。

[0129] 向量（又称矢量）是一个既有长度又有方向的量，在一个空间坐标系中，可以用坐标系的各个维度上的分量去描述，比如在二维直角坐标系中，从原点到 (3,4) 这一点的向量就是一个长度为 5、方向为原点到 (3,4) 这一点的向量，这个向量就可以表示为 (3,4)，即该向量在 x 这个维度上的分量为 3，在 y 这个维度上的分量为 4；由于计算机无法直接对自然语言的文本做出任何理解或处理工作，因此，本发明对问题和问答实例进行向量化，即仅保留问题和问答实例中的若干关键词，这些关键词以高维向量的形式存在，以便于进行计算；一个关键词就是问题向量或实例向量的一个维度，相当于直角坐标系的 x 维度或 y 维度，由于问题或问答实例中包含很多不同的关键词，所以问题向量和实例向量均是一个高维向量。

[0130] 二元字符串指相邻两个单字组合而成的字符串，比如“问答实例”包含的二元字符串为：问答、答实和实例。

[0131] 相似度计算是指计算两个向量的相似程度，即将该两个向量看作是高维空间的两个点的相近程度；余弦相似度就是计算两个向量之间的夹角大小，以此来衡量两个向量的相近程度。

[0132] 加权投票的每一票计数不是简单的 1 票，而是与投票方的权重有关，权重越大的，其投一次票得到的计数就越大。

[0133] 图 2 为本发明构建问答实例库的方法流程图。现结合图 2，对本发明构建问答实例库的方法进行说明，具体如下：

[0134] 步骤 20：确定需自动问答的知识点，为需自动问答的知识点分配问题 ID；

[0135] 该步骤提及的需自动问答的知识点为需要采用非人工服务的方式进行回复的知识点范围，该知识点范围可根据该问答实例库所属的领域进行设定；或者从知识点文案库中选定需自动问答的知识点的范围。

[0136] 其中，知识点文案库可为人工答复时提供知识点回答文案的数据库，该数据库中的每个知识点有唯一的 ID、知识点描述及知识点回答文案。

[0137] 该步骤中提及为需自动问答的知识点分配的问题 ID 也是唯一的。

[0138] 步骤 21：获得与需自动问答的知识点对应的问答实例并分配回答知识点 ID；

[0139] 该步骤包括：根据需自动问答的知识点，对人工回答记录进行采样，获得与需自动问答的知识点对应的问答实例，为问答实例包含的知识点分配回答知识点 ID。

[0140] 其中，为问答实例包含的知识点分配的回答知识点 ID 可以参考知识点文案库中的知识点 ID，比如，将相同知识点对应的回答知识点 ID 及知识点文案库中的知识点 ID 建立

对应关系，或者将相同知识点对应的回答知识点 ID 及知识点文案库中的知识点 ID 设为相同的内容。

[0141] 其中，对人工回答记录进行采样，获得与需自动问答的知识点对应的问答实例包括：确定进行采样的知识点的样本数量 n；进行采样的知识点为需自动回答的知识点；n 为自然数；从人工回答记录中选择包含知识点的 n 个问题实例；计算 n 个问题实例的问题的整体长度方差；判断问题的整体长度方差是否低于第一阈值，如果是，则将 n 个问题实例作为选择的 n 个包含需自动回答的知识点的问答实例，否则，去除 n 个问题实例中问题长度与 n 个问题实例的长度平均值的差值最大的一个问题实例，从人工回答记录中再选择一个包含知识点的问题实例，执行计算 n 个问题实例的问题的整体长度方差的步骤。

[0142] 步骤计算 n 个问题实例的问题的整体长度方差，可利用$\frac{1}{n}\sum_{q_i \in K_j}\left(len(q_i)-E_{K_j}(len)\right)^2$计算选为采样样本的 n 个问题实例的问题的整体长度方差；上述公式中，q_i 为知识点 K_j 的问题样本，len(q_i) 为 q_i 包含的词数量，$E_{K_j}(len)$为知识点 K_j 中所有问题长度的平均值；$len(q_i)-E_{K_j}(len)$为问题长度与 n 个问题实例的长度平均值的差值。

[0143] 上述步骤中，若计算获得的问题的整体长度方差大于预设的第一阈值，则去掉$len(q_i)-E_{K_j}(len)$绝对值较大的 q_i，即长度与平均值相差较大的问题实例，然后采样新的问题实例进行补充，再进行上述问题的整体长度方差的计算过程，直至满足第一阈值的要求。

[0144] 步骤 22：向量化问答实例包含的问题，获得问题向量；

[0145] 该步骤包括：步骤 a，提取问答实例中问题的关键词、二元字符串及特殊词性；步骤 b，将关键词作为向量元素，计算每一向量元素的权重；步骤 c，利用每个向量元素的权重计算问答实例的向量的长度；步骤 d，将向量元素、向量元素的权重及向量的长度作为问题向量。

[0146] 其中，步骤 b 可利用$ch(i)=\frac{N*[A*N-CF*TF]^2}{CF*TF*(N-CF)*(N-TF)}$计算获得向量元素与知识点之间的卡方值 ch(i)，或者利用现有的卡方值计算公式计算向量元素与知识点之间的卡方值 ch(i)；再利用卡方值 ch(i) 及 weight(c,a)=ln(chi+b) 计算获得预设的知识点范围内的每一个向量元素的权重。上述公式中，N 为样本总数量，CF 为每个知识点的样本数量，TF 为每个向量元素出现的样本数量，A 为向量元素和知识点共同出现的样本数量，ch(i) 为向量元素 c 与知识点 a 的卡方值，weight(c,a) 表示向量元素 c 在知识点 a 中的权重，b 是平滑值，b 可取小数，比如可取 0.5。

[0147] 步骤 c 中可利用$|v|=\sqrt{\sum_i^n w_i * w_i}$计算问答实例的向量长度；上述公式中，|v| 为向量长度，w_i 为向量问答实例中每个向量元素的权重；n 为选择的问答实例的数量。

[0148] 步骤 23：将问答实例以三元组的形式进行存储；

[0149] 本发明所构建的问答实例库中任一问答实例是以三元组的形式进行存储的，该三元组包含问题 ID、问题向量及回答知识点 ID，具体形式可为：

[0150] <问题 ID，问题向量（元素、权重、长度），回答知识点 ID>。

[0151] 比如：采样后编号为 1500456 的问题为“我想关注别人！”，在人工客服回答历史记

录中是采用知识点文案库中 ID 为 15 的知识点文案进行回答的，那么这个问答实例经过本发明的上述处理后在问答实例库中存储形式如下表所示：

[0152]

问题 ID	问题向量	回答知识点
1500456	[我想 :2.1 关注 :4.6]/5.05	15

[0153] 表一

[0154] 表一中的问题向量中有两个元素：我想和关注，如果实例问题文本中还有数词词性和时间词词性，那么这两者也是向量元素，每个向量元素后面用“:”分隔的数值（2.1、4.6）就是这个元素在 ID 为 15 的知识点中的权重，5.05 就是这个向量的长度；在构建问答实例库时计算出向量长度，能够提高问答系统在回复时的响应效率。

[0155] 图 3 为本发明自动问答方法的流程图。现结合图 3，对本发明自动问答方法进行说明，具体如下：

[0156] 步骤 30：将接收到的问题向量化获得问题向量；

[0157] 该步骤中的问题向量包含多个向量元素。

[0158] 该步骤包括：将接收到的问题规整、分词和归一化处理，提取关键词、二元字符串和词性；将关键词作为向量元素；将向量元素、向量元素的权重及向量的长度作为问题向量。

[0159] 步骤将接收到的问题规整、分词和归一化处理，提取关键词、二元字符串和词性中，可采用现有的方法对问题进行规整、分词和归一化处理，在此不再对所采用的详细的处理方法进行赘述。

[0160] 本发明的向量元素以关键词为主，二元字符串作为对关键词不足时的补充，词性是某些特定的词性，不需要具体的词，只需要保存其词性即可。

[0161] 该步骤中向量元素包含的向量元素的权重及向量的长度被设置为空。

[0162] 步骤 31：根据向量元素检索问答实例库，获得多个实例向量；

[0163] 该步骤中的多个实例向量的任一实例向量至少包含一个向量元素。

[0164] 该步骤中，将向量元素中包含的关键词作为进行检索的知识点，从问答实例库中检索获得包含相同知识点的实例向量。

[0165] 步骤 32：利用相似度计算公式，计算问题向量与多个实例向量的相似度；

[0166] 该步骤包括：步骤 e，将多个实例向量按照其包含的回答知识点 ID 进行归类；步骤 f，对于同一回答知识点 ID，动态调整问题向量包含的向量元素在相同的回答知识点 ID 对应的实例向量中的权重，获得调整后的向量元素；步骤 g，利用调整后的向量元素在相同回答知识点 ID 对应的实例向量中的权重，计算调整后的向量元素与实例向量的余弦相似度。

[0167] 步骤将多个实例向量按照其包含的回答知识点 ID 进行归类中，检索得到的所有实例向量可按照其所属“回答知识点 ID”进行归类，即“回答知识点 ID”相同的实例向量放在一起，具体可参见图 4 所示，这样做是能够使得问题向量 v_q 与同一个回答知识点 ID 下的所有实例向量可以在同一批完成计算，提高工作效率。

[0168] 步骤 f 中，动态调整的方法可以为：问题向量 v_q 在与回答知识点 ID1 下所有实例进行相似度计算时，如果 v_q 中的元素在问答实例库中“回答知识点 ID 为 ID1”的实例中出

现过，则其权重就取值为它在 ID1 中的权重，否则，其权重取默认值，该默认值可根据需要进行设置，比如设置为问答实例库中回答知识点 ID 为 ID1 的实例向量中向量元素的权重的最大值的 80%。

[0169] 其中，每次动态调整权重之后，就计算 v_q 与当前 ID 下每一个问答实例向量 v_c 的余弦相似度。

[0170] 步骤利用调整后的向量元素在相同回答知识点 ID 对应的实例向量中的权重，计算调整后的向量元素与实例向量的余弦相似度中，可利用 $sim(v_q, v_c) = \frac{\sum_i w_i^{(q)} * w_i^{(c)}}{|v_q| * |v_c|}$ 计算调整后的向量元素所在的问题向量与实例向量的余弦相似度。

[0171] 其中，v_q 为问题向量；v_c 为实例向量；$|v_q| * |v_c|$ 表示问题向量的长度与实例向量的长度的乘积；$w_i^{(q)}$ 和 $w_i^{(c)}$ 表示两个向量中相同向量元素对应的权重。

[0172] 步骤 33：利用相似度确定答复知识点，输出与答复知识点对应的数据。

[0173] 该步骤包括：步骤 h，将相似度转换为实例向量与问题向量之间的距离；步骤 i，将距离小于第二阈值的实例向量作为候选实例向量；步骤 j，利用候选实例向量计算获得其对应的回答知识点 ID 的加权票数；步骤 k，在回答知识点 ID 的平均票数大于第三阈值时或在为回答知识点 ID 投票的实例向量的数量大于第四阈值时，将回答知识点 ID 确定为候选知识点 ID；步骤 l，选择加权票数排列在前 L 位的候选知识点 ID 对应的知识点为答复知识点；步骤 m，利用候选知识点 ID 从知识点文案库中读取排列在前 L 位的回答内容，并按照加权票数的排列顺序输出 L 个回答内容。

[0174] 其中，L 为自然数。

[0175] 步骤 h 中，可利用 $dis(v_c, v_q) = \log(\frac{1}{sim(v_c, v_q)})$ 及 $sim(v_c, v_q) > \lambda > 0$ 计算实例向量与问题向量之间的距离；具体地，忽略 $sim(v_c, v_q)$ 低于 λ 的实例向量，将保留的实例向量按照距离计算公式 $dis(v_c, v_q) = \log(\frac{1}{sim(v_c, v_q)})$ 进行计算。上述公式中，$sim(v_c, v_q)$ 为实例向量与问题向量的余弦相似度；λ 为相似度阈值。

[0176] 步骤 j 中，可利用 $vote(ID_j) = \sum_{v_i \in ID_j}^{m} \frac{1}{dis < v_i, v_q >^2}$ 计算候选实例向量对应的回答知识点 ID 的加权票数；上述公式中，ID_j 为实例向量包含的回答知识点 ID；v_i 为属于回答知识点 ID_j 的实例向量；v_q 是问题向量；$dis<v_i, v_q>$ 为实例向量与问题向量之间的距离；m 为属于回答知识点 ID_j 的实例向量的数量。

[0177] 步骤 j 中，每一个回答知识点 ID 所得总票数由候选实例中属于该知识点的实例向量加权求和而来，比如图 4 中回答知识点 ID1 所得票数就是由实例 id11、id12 等为其投票。

[0178] 步骤 k 中回答知识点 ID 的平均票数为回答知识点 ID 加权票数除以属于回答知识点 ID 的实例向量的数量获得的商。

[0179] 步骤 l 及步骤 m 中，在 L 取 1 时，按照票数降序排列之后，挑选排名第一的回答知识点 ID，从知识点文案库中读取相同 ID 的知识点对应的知识点回答文案，作为对用户提交的问题的回答内容，可进一步讲排名第二的回答知识点 ID 对应的知识点文案库中的回答文案作为候选回答反馈给提交问题的用户。

[0180] 优选地，步骤 30 之前还包括如图 2 所示构建问答实例库的方法，在此不再赘述具体的内容，可参见步骤 20 至步骤 23 的内容。

[0181] 优选地，步骤 33 之后还包括：如果不存在符合的答复知识点，则将问题转发至人工进行答复。

[0182] 现举一具体实例，对本发明的方法进行说明：以微博客服为例，用户向自助客服提交的问题为："我想关注别人，没问题吧？谢谢。"下面详细描述利用本发明的方法输出回答内容的过程。

[0183] 问题规整，去掉无关字符串，如"谢谢"，自动标注出所属领域的关键词或与产品有关的词，如"关注"；对问题进行分词得到：我 \ 想 \ 关注 \ 别人 \ 没 \ 问题；提取关键词"关注"，并提取由相邻单字词（单字词就是分词后只含一个汉字的词）组成的二元字符串作为关键词补充，比如将"我"和"想"组成"我想"作为补充，二元字符串作为向量元素需要满足一定条件，比如关键词数量较少（低于设定阈值）时补充为向量元素；如果实例问题文本中还有数词词性和时间词词性，那么也可将其作为向量元素；生成问题向量，即［我想：?关注：?问题：?］，其中，？表示权重待定，需要在计算过程中动态调整；检索问答实例库，即利用"我想"和"关注"去检索问答实例库，得到包含问题向量中至少一个向量元素的所有实例向量列表，"问题"一词没有检索到任何实例，按照回答知识点 ID 分类，如表二所示：

[0184]

问题 ID	问题向量	回答知识点
1500456	［我想：2.1 关注：4.6］/5.05	15
1500457	［无法：2.1 关注：4.6］/5.05	15
1500458	［怎么：2.1 关注：4.6］/5.05	15
1500459	［我想：2.0 认证：4.7］/5.10	16
1500459	［我想：2.0 认证：4.7］/5.10	16

[0185]

[0186] 表二

[0187] 按照回答知识点的 ID 分批计算问题向量与检索到的实例向量之间的相似性，即将问题向量分别与回答知识点 ID=15 对应的实例向量和回答知识点 ID=16 对应的实例向量计算相似性；

[0188] 具体地，计算问题向量［我想：?关注：?问题：?］与回答知识点 ID 为 15 下的实例向量 1500456、1500457、1500458 之间的相似性，先将问题向量［我想：?关注：?问题：?］中向量元素权重根据回答知识点 ID=15 中的相应向量元素的权重进行调整，得到调整后的问题向量［我想：2.1 关注：4.6 问题：3.68］/6.24，然后分别与 1500456、1500457、1500458 实例向量计算余弦相似度，分别为 0.80、0.67、0.67。

[0189] 将这三个相似度分别转换为距离值 0.22、0.40、0.40，若第二阈值取值为 0.91，则三个实例向量均可以参与投票；这三个实例向量均为其所属回答知识点 ID=15 投票，

所投票数分别为 20.66、6.25、6.25，回答知识点 ID=15 最终得票为 33.16，其平均票数为 (33.16)/3=11.05，且为其投票的实例向量为 3；问题向量再与回答知识点 ID=16 下的实例向量进行前述处理后，回答知识点 ID=16 也得到了相应的相似度，两个实例均为 0.12，转换为距离值就是 2.12，大于第二阈值，不能参与投票；取排名第一的回答知识点 ID=15 的知识点回答文案，输出作为对用户的回答。

[0190] 图 5 为发明自动问答系统的结构示意图。现结合图 5，对本发明自动问答系统的结构进行说明，具体如下：

[0191] 本发明的自动问答系统包含：问答实例库 50、问题解析模块 51 及答案生成模块 52。

[0192] 问答实例库 50 以三元组的形式存储问答实例。其中，任一问答实例的三元组包含问题 ID、问题向量及回答知识点 ID，将以三元组形式存储的问答实例称为实例向量。

[0193] 问题解析模块 51 将接收到的问题向量化获得问题向量，根据向量元素检索问答实例库 50，获得多个实例向量，并输出多个实例向量至答案生成模块 52。其中，任一实例向量至少包含一个向量元素，问题向量包含多个向量元素。

[0194] 答案生成模块 52 利用相似度计算公式，计算问题向量与多个实例向量的相似度，利用相似度确定答复知识点，输出与答复知识点对应的数据。

[0195] 其中，问题解析模块 51 包含：第一向量化单元 511 和检索单元 512。

[0196] 第一向量化单元 511 将接收到的问题规整、分词和归一化处理，提取关键词、二元字符串和词性，将关键词作为向量元素，将向量元素、向量元素的权重及向量的长度作为问题向量。其中，向量元素的权重和向量长度被设置为空。

[0197] 检索单元 512 根据向量元素检索问答实例库 50，获得多个实例向量，并输出多个实例向量至答案生成模块 52。

[0198] 其中，答案生成模块 52 包括：实例挑选单元 521、知识点挑选单元 522 和答案筛选单元 523。

[0199] 实例挑选单元 521 将多个实例向量按照其包含的回答知识点 ID 进行归类，对于同一回答知识点 ID，动态调整问题向量包含的向量元素在相同的回答知识点 ID 对应的实例向量中的权重，获得调整后的向量元素。

[0200] 实例挑选单元 521 利用调整后的向量元素在相同回答知识点 ID 对应的实例向量中的权重，计算调整后的向量元素与实例向量的距离，将距离小于第二阈值的实例向量作为候选实例向量并输出至知识点挑选单元 522。

[0201] 知识点挑选单元 522 利用候选实例向量计算获得其对应的回答知识点 ID 的加权票数，在回答知识点 ID 的平均票数大于第三阈值时或在为回答知识点 ID 投票的实例向量的数量大于第四阈值时，将回答知识点 ID 确定为候选知识点 ID 并输出至答案筛选单元 523。

[0202] 答案筛选单元 523 选择加权票数排列在前 L 位的候选知识点 ID 对应的知识点为答复知识点，利用候选知识点 ID 从知识点文案库中读取排列在前 L 位的回答内容，并按照加权票数的排列顺序输出 L 个回答内容。其中，L 为自然数。

[0203] 优选地，本发明的自动问答系统还可与保存了人工回复的回答记录的数据库建立连接，以利用人工回复的回答记录构建问答实例库。本发明的自动问答系统还包含：构建模

块 53。构建模块 53 采样人工回答记录并向量化，生成问答实例库。

[0204] 其中，构建模块 53 包含：知识点确定单元 531、采样单元 532 和第二向量化单元 533。

[0205] 知识点确定单元 531 确定需自动问答的知识点，为需自动问答的知识点分配问题 ID，输出需自动问答的知识点及其对应的问题 ID 至采样单元 532。

[0206] 采样单元 532 根据需自动问答的知识点，对人工回答记录进行采样，获得与需自动问答的知识点对应的问答实例，为问答实例包含的回答数据分配回答知识点 ID，输出问题 ID、回答知识点 ID 及问答实例至第二向量化单元 533。

[0207] 第二向量化单元 533 向量化问答实例包含的问题，获得问题向量，将问答实例以三元组的形式存储于问答实例库 50 中。

[0208] 优选地，该系统还可包含：知识点文案库 54。知识点文案库 54 以三元组形式保存知识点向量；任一知识点向量的三元组包含回答知识点 ID、知识点描述及知识点回答文案。

[0209] 本发明的上述较佳实施例中，由于问答实例库构建是自动从记录有回答记录的数据库中采样，所以不需要客服人员持续进行问答实例库的维护，仅仅需要不太经常的领域知识和行业知识的更新；由于回答知识点 ID 和知识点回答文案分离，且知识点回答文案不参与计算过程，所以知识点回答文案可修改，且修改知识点回答文案完全不影响自动问答系统的工作；由于处理问题均采样自人工客服的问答记录，所以自助客服的问答和人工客服无异，且替代人工客服的工作量大大增加，提高了工作效率；由于答案产生过程采用了实例加权投票方式，所以给出答案可信度高；由于计算步骤简单，无需模板匹配，提高了响应时间，降低了成本，提高了工作效率。

[0210] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序可以存储于一计算机可读取存储介质中，如：ROM/RAM、磁碟、光盘等。

[0211] 以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以作出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

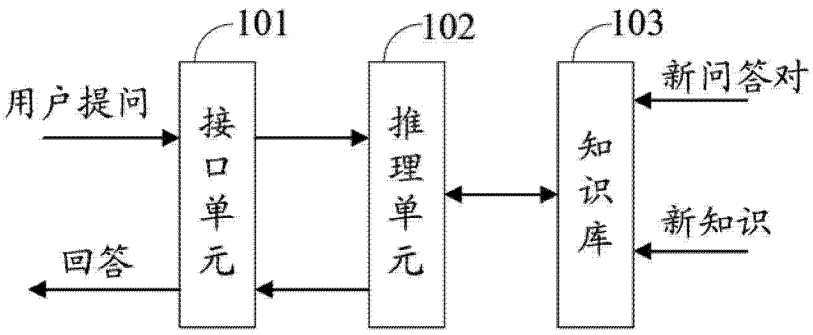

101
102
103
用户提问
接口单元
推理单元
知识库
新问答对
新知识
回答

图 1

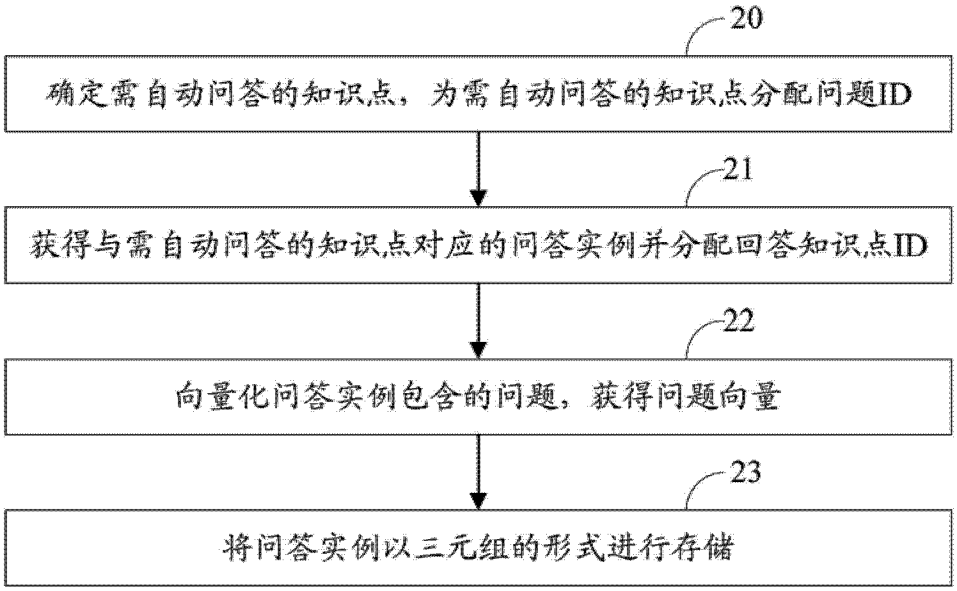

20
确定需自动问答的知识点，为需自动问答的知识点分配问题ID
21
获得与需自动问答的知识点对应的问答实例并分配回答知识点ID
22
向量化问答实例包含的问题，获得问题向量
23
将问答实例以三元组的形式进行存储

图 2

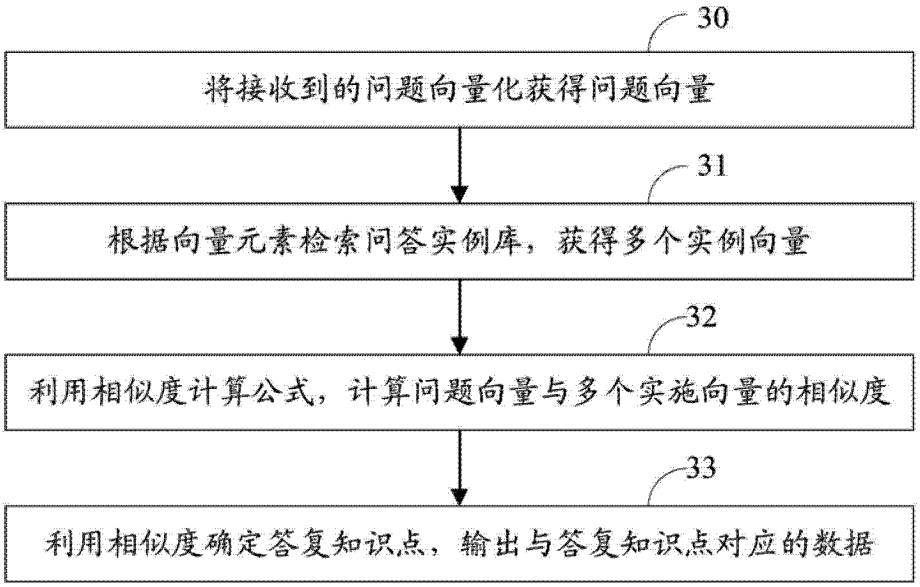

30
将接收到的问题向量化获得问题向量
31
根据向量元素检索问答实例库，获得多个实例向量
32
利用相似度计算公式，计算问题向量与多个实施向量的相似度
33
利用相似度确定答复知识点，输出与答复知识点对应的数据

图 3

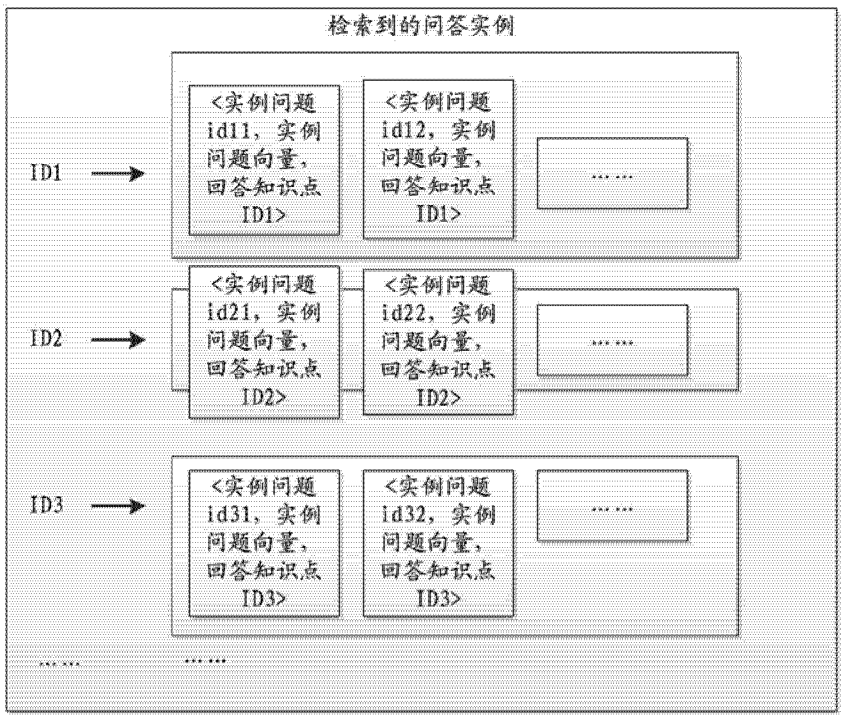

检索到的问答实例
ID1
<实例问题 id11，实例问题向量，回答知识点 ID1>
<实例问题 id12，实例问题向量，回答知识点 ID1>
……
ID2
<实例问题 id21，实例问题向量，回答知识点 ID2>
<实例问题 id22，实例问题向量，回答知识点 ID2>
……
ID3
<实例问题 id31，实例问题向量，回答知识点 ID3>
<实例问题 id32，实例问题向量，回答知识点 ID3>
……
……
……

图 4

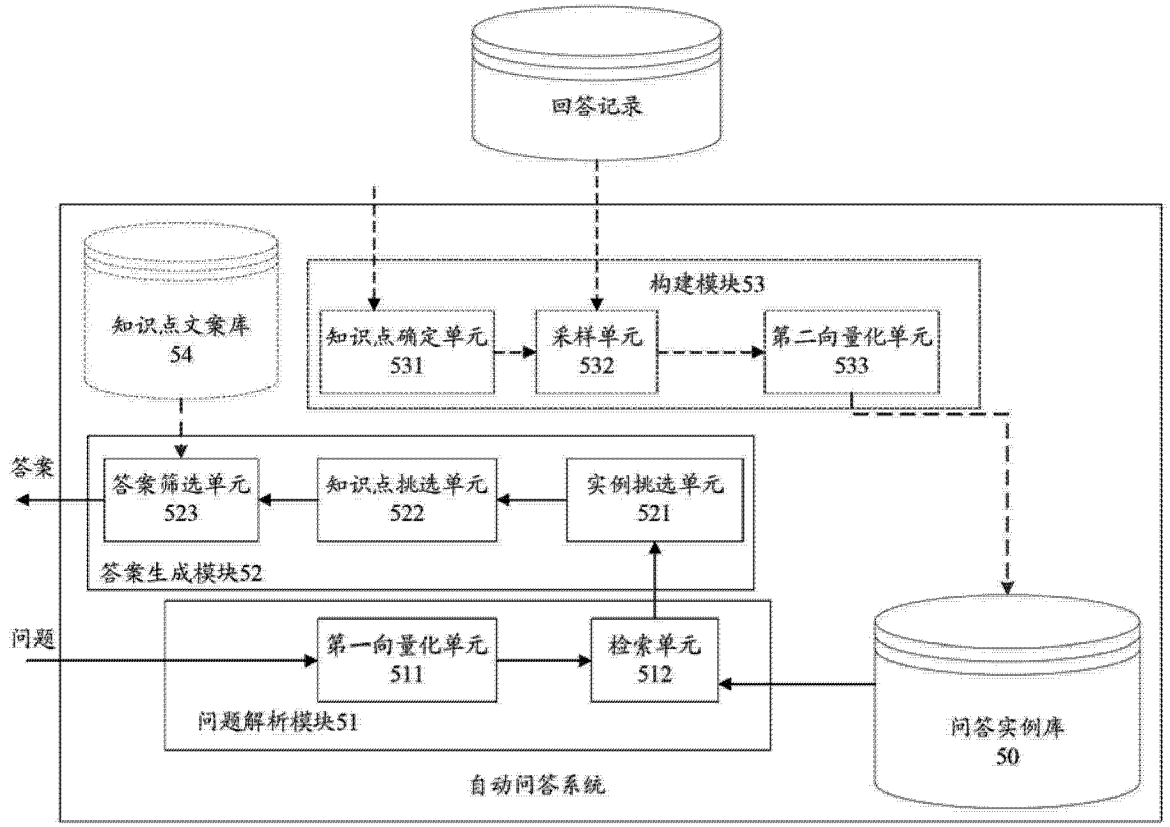

回答记录
知识点文案库
54
构建模块53
知识点确定单元
531
采样单元
532
第二向量化单元
533
答案
答案筛选单元
523
知识点挑选单元
522
实例挑选单元
521
答案生成模块52
问题
第一向量化单元
511
检索单元
512
问题解析模块51
问答实例库
50
自动问答系统

图 5